LUDWIG MAXIMILIANS-UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR MATHEMATIK, INFORMATIK UND STATISTIK

INSTITUT FÜR STATISTIK

BACHELORARBEIT

WISSENSCHAFTLICHE ARBEIT ZUR ERLANGUNG DES AKADEMISCHEN GRADES
BACHELOR OF SCIENCE

DAS COMPLEMENTARY-LOG-LOG MODELL ZUR SCHÄTZUNG VON ROC-KURVEN DIAGNOSTISCHER STUDIEN

Ruben Camilo Wißkott

Matrikelnr. 11745645

betreut durch

Prof. Dr. Annika Hoyer
Institut für Biometrie und Epidemiologie
Deutsches Diabetes-Zentrum Düsseldorf

20. März 2020

Zusammenfassung

In Diagnosestudien werden zumeist nur gruppierte Daten in Form von 4-Felder-Tafeln zu ausgewählten Schwellenwerten berichtet. Dies führt zum Problem, dass auf Grundlage dieser Daten keine parametrische Receiver Operating Characteristic (ROC)-Kurve mittels des bekannten ROC-GLM-Modells geschätzt werden kann. Eine Möglichkeit, um dem Problem entgegen zu kommen, ist es gruppierte Daten als intervall-zensiert aufzufassen und Methoden aus der Survival Analysis darauf anzuwenden. Hierbei eignet sich das complementary-log-log Modell zur parametrischen Schätzung der Spezifität und Sensitivität. Der größte Aufwand liegt in der geeigneten Datenaufbereitung, hierfür können jedoch vorgefertigte Algorithmen eine Abhilfe schaffen.

Inhaltsverzeichnis

1. Problemstellung

In diesem Kapitel möchte ich die Problemstellung, welche mir freundlicherweise von Prof. Dr. Annika Hoyer bereitgestellt wurde, kurz durchgehen.

Diese Arbeit beschäftigt sich mit dem Thema *"Das Complementary-log-log Modell zur Schätzung von ROC-Kurven diagnostischer Studien"*. Dabei möchten wir einen neuen Modellansatz diskutieren um aus Diagnosestudien eine Receiver Operating Characteristic (ROC)-Kurve schätzen zu können. Diagnosestudien stellen eine eigenständige Form von medizinischen Studien dar, bei denen das Ziel verfolgt wird, neue Verfahren zur Diagnose von Erkrankungen zu evaluieren. In solchen Studien wird üblicherweise ein neuer diagnostischer Test mit dem bisherigen Standardverfahren, dem sogenannten Goldstandard, hinsichtlich der Kennzahlen *Sensitivität* und *Spezifität* verglichen. Bei der Sensitivität handelt es sich dabei um die Wahrscheinlichkeit, dass der neue Test positiv ausfällt, wenn bei dem Individuum tatsächlich eine Krankheit vorliegt. Die Spezifität beschreibt die Wahrscheinlichkeit, dass der neue Test negativ ausfällt, wenn die Erkrankung nicht vorhanden ist. Unsere Datengrundlage stellen sogenannte 4-Felder-Tafeln dar, welche die Anzahl an richtig-positiven (True Positive, TP), richtig-negativen (True Negative, TN), falsch-positiven (False Positive, FP) und falsch-negativen (False Negative, FN) Testergebnissen enthalten.

In der Epidemiologie und der biomedizinischen Forschung werden die Werte von einem stetigen diagnostischem Biomarker erhoben und mittels mehrerer geeigneter Schwellenwerte verglichen. Überschreitet der Biomarkerwert einen ausgewählten Schwellenwert, so wird das Individuum als krank klassifiziert. Dies bedeutet, dass für jeden untersuchten Schwellenwert Daten aus einer zugehörigen 4-Felder-Tafel vorliegen. Die zugehörigen Kennzahlen - Sensitivität und Spezifität - werden über ROC-Kurven grafisch veranschaulicht. Eine etablierte Möglichkeit zur Schätzung solcher ROC-Kurven stellt der ROC-GLM-Schätzer dar, der auf einem Probit-Modell basiert (Pepe, 2003, Kapitel 5). Die Publikation von Hoyer et al. (2017) konnte zeigen, dass die Daten von Diagnosestudien als intervall-zensiert angesehen und mithilfe von Verfahren aus der Survival-Analyse ausgewertet werden können.

Im Rahmen dieser Arbeit wird eine Einführung und Durchführung der Auswertung intervallzensierter Daten mittels des complementary-log-log Modells vorgestellt. Dies ist ein generalisiertes lineares Regressionsmodell mit einer binären Zielvariable und einer complementary-log-log Linkfunktion. Zum Abschluss wird ein kurzer visueller Vergleich der Standardmethode des ROC-GLM-Schätzers mit unserem complementary-log-log Modell vorgestellt. Da der ROC-GLM-Schätzer auf den Individualdaten und das complementary-log-log Modell auf intervall-zensierten Daten beruhen, werden wir die Individualdaten auf gruppierte Daten aggregieren um das complementary-log-log Modell anwenden zu können und somit beide Ansätze vergleichbar zu machen.

Anmerkungen

Der Großteil der aufgebrachten Zeit wurde durch die Erstellung der Funktionen und Algorithmen mittels der Programmiersprache R für die individuelle Datenaufbereitung beansprucht. Alle selbsterstellten Funktionen und Algorithmen sind im Anhang A zu finden.

Für das Verständnis bedarf es eine Einarbeitung in die Theorie der diskreten Survival-Analyse, hierbei hatte sich das Werk von Tutz and Schmid (2016) als sehr hilfreich herausgestellt und hat somit eine Grundlage für diese Arbeit gelegt.

Weiterhin möchte ich Prof. Dr. Annika Hoyer vom Institut für Biometrie und Epidemiologie des Deutschen Diabetes-Zentrums in Düsseldorf für die Vergabe der Problemstellung, die bereitgestellten Daten sowie für die zuverlässige Betreuung danken.

2. Grundlagen

In diesem Kapitel werden wir die Grundlagen sowie Grundbegriffe, welche als Voraussetzung für die in den kommenden Kapiteln verwendeten Regressionsmodelle dienen, kurz einführen und erläutern. Dafür betrachten wir das Konzept von bedingten Wahrscheinlichkeiten, sowie die für die Datengrundlage verwendeten diagnostischen Testverfahren.

2.1 Bedingte Wahrscheinlichkeiten

Im späteren Verlauf werden wir auf Wahrscheinlichkeitsbegriffe stoßen, welche das Verständnis von *bedingten Wahrscheinlichkeiten* voraussetzen.

Definition 1 (Bedingte Wahrscheinlichkeit)**.**
Seien A und B beliebige Ereignisse und es gilt $P(B) > 0$, dann bezeichnen wir mit

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

die Wahrscheinlichkeit für das Eintreten von A unter der Bedingung B. Mit $P(A \cap B)$ wird die Wahrscheinlichkeit bezeichnet, dass die Ereignisse A und B gemeinsam auftreten.

Eine häufige Anwendung der bedingten Wahrscheinlichkeit führt uns zum *Satz von Bayes*, welcher eine der wichtigsten Grundlagen für zahlreiche Anwendungen in der Statistik darstellt und daher kurz erwähnt wird. Haben wir die bedingte Wahrscheinlichkeit $P(A|B)$ gegeben, aber interessieren wir uns eigentlich für die Wahrscheinlichkeit des Ereignisses B gegeben A, so liefert folgender Satz eine Antwort darauf.

Resultat 1 (Satz von Bayes)**.**
Seien die bedingte Wahrscheinlichkeit $P(A|B)$ sowie die Wahrscheinlichkeiten $P(A)$ und $P(B)$ bekannt. Dann gilt

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A)}$$

Die Wahrscheinlichkeit $P(B)$ wird auch als a-priori Wahrscheinlichkeit bezeichnet. Mit $P(B|A)$ bezeichnen wir somit die a-posteriori Wahrscheinlichkeit für das Ereignis B.

2.2 Diagnostische Testverfahren

Nachdem wir die benötigten Wahrscheinlichkeitsbegriffe betrachtet haben, widmen wir uns es jetzt den Grundlagen diagnostischer Testverfahren nach Pepe (2003).

Wir bezeichnen mit der binärkodierten Variable K, ob eine Krankheit bei einem Individuum tatsächlich vorliegt oder nicht. Mit $K = 1$ wird eine tatsächlich vorliegende Krankheit und mit $K = 0$ eine tatsächlich nicht vorliegende Krankheit bezeichnet:

$$K = \begin{cases} 1 & \text{Krankheit liegt vor} \\ 0 & \text{Krankheit liegt nicht vor} \end{cases}$$

Die binäre Variable T hingegen bezeichnet das vorliegende diagnostische Testergebnis, auch *Klassifizierung* genannt. Mit $T = 1$ wird ein positives Testergebnis und mit $T = 0$ ein negatives Testergebnis bezeichnet:

$$T = \begin{cases} 1 & \text{positives Testergebnis} \\ 0 & \text{negatives Testergebnis} \end{cases}$$

Haben wir nun für jedes einzelne Individuum Testergebnisse T vorliegen und kennen wir zu jedem Testergebniss auch das Vorhandensein der tatsächlichen Krankheit K, so können wir uns Folgendes veranschaulichen. Jedes Testergebnis können wir einer der vier Kategorien - *True Positive, True Negative, False Negative* und *False Positive* - zuordnen.

Wie man aus der untenstehenden Tafel ablesen kann, bedeutet beispielsweise die Zugehörigkeit eines Individuums zur Kategorie *True Positive*, dass es positiv klassifiziert wurde und bei dem auch eine tatsächliche Krankheit vorliegt. Für Kategorie *False Negative* bedeutet dies, dass ein Individuum ein negatives Testergebnis erhalten hat, obwohl eine tatsächliche Krankheit vorhanden ist.

	$K = 1$	$K = 0$
$T = 1$	True Positive	False Positive
$T = 0$	False Negative	True Negative

Definition 2 (Vier-Felder Tafel)**.**
Seien K und T zwei binärkodierte Variablen sowie a, b, c und d Anzahlen von Individuen, die entsprechend als True Positive, False Positive, False Negative und True Negative kategorisiert wurden und N bezeichne die Gesamtanzahl aller Individuen. Dann definieren wir folgende Tabelle als eine Vier-Felder Tafel, auch (2×2)-Kontingenztafel genannt

	$K = 1$	$K = 0$	
$T = 1$	a	b	$a + b$
$T = 0$	c	d	$c + d$
	$a + c$	$b + d$	N

Im Kommenden betrachten wir wichtige Kennzahlen, welche aus der *Vier-Felder Tafel* hergeleitet werden können.

Definition 3 (False-Positive-Fraction, True-Positive-Fraction)**.**
Wir bezeichnen mit

$$FPF = P(T = 1|K = 0)$$

die False-Positive-Fraction, also die Wahrscheinlichkeit für $T = 1$ unter der Bedingung $K = 0$ sowie

$$TPF = P(T = 1|K = 1)$$

die True-Positive-Fraction, also die Wahrscheinlichkeit für $T = 1$ unter der Bedingung $K = 1$.

Analog werden zwei weitere komplementäre *Fractions*, nämlich die *True-Negative-Fraction*

$$TNF = P(T = 0|K = 0)$$

und die *False-Negative-Fraction*

$$FNF = P(T = 0|K = 1)$$

dafiniert. Dabei gelten folgende Zusammenhänge

$$1 = FPF + TPF, \qquad 1 = TNF + FNF,$$

die mittels *bedingter Wahrscheinlichkeiten* für komplemetäre Ereignisse hergeleitet werden können und im weiteren Verlauf dieser Arbeit verwendet werden.

Resultat 2 (Schätzung von FPF und TPF)**.**
Seien a, b, c und d wie in Definition 3. Dann ist ein Schätzer für die FPF gegeben durch $\frac{b}{b+d}$. Analog ist $\frac{a}{a+c}$ ein Schätzer für die TPF.

Begründung. Durch Anwendung von bedingten Wahrscheinlichkeiten erhalten wir

$$FPF = P(T = 1|K = 0) = \frac{P(T = 1, K = 0)}{P(K = 0)}.$$

Hierbei wird die gemeinsame Wahrscheinlichkeit $P(T = 1, K = 0)$ durch die relative Häufigkeit $\frac{b}{N}$ und $P(K = 0)$ durch $\frac{b+d}{N}$ geschätzt. Analog verfahren wir mit TPF. □

Im weiteren Verlauf dieser Arbeit machen wir keine notationelle Unterscheidung zwischen einem theoretischen und einem geschätzten Wert, da dieser Unterschied aus dem Kontext klar sein wird.

Definition 4 (Prävalenz)**.**
Seien weiterhin a, b, c und d wie in Definition 3 und sei $n_{(K=1)} = a + c$ die Anzahl der Individuen mit einer bestimmten Krankheit K innerhalb einer Population mit insgesamt N Individuen. Dann bezeichnet

$$P(K = 1) = \frac{n_{(K=1)}}{N} = \frac{a + c}{N}$$

die (geschätzte) Prävalenz für eine bestimmte Krankheit.

Die Prävalenz kann man als a-priori Wahrscheinlichkeit, wie sie im *Satzes von Bayes* definiert wurde, verwenden.

Nun definieren wir zwei weitere Kennzahlen, die als Gütekriterien für diagnostische Tests verwendet werden - die *Sensitivität* und die *Spezifität*. Um einen Zusammenhang zwischen diesen Begriffen und den zuletzt eingeführten *Fractions* zu erschaffen, benutzen im Weiteren a, b, c und d wie in Definition 3.

Definition 5 (Sensitivität)**.**
Sei $n_{(T=1|K=1)}$ die Anzahl der Personen mit einem positiven Testergebnis T und einer tatsächlich vorliegenden Krankheit K. Des weiteren sei $n_{(K=1)}$ die Anzahl der Personen mit einer bestimmten Krankheit K. So bezeichnen wir mit

$$P(T = 1|K = 1) = \frac{n_{(T=1|K=1)}}{n_{(K=1)}} = \frac{a}{a + c}$$

die (geschätzte) Sensitivität eines Tests.

Dabei können wir sehen, dass die *Sensitivität* eines Tests exakt der *True-Positive-Fraction (TPF)* entspricht und nur einen weiteren in der Epidemiologie und Medizin häufig verwendeten Begriff darstellt. Mit der Sensitivität schätzen wir somit die bedingte Wahrscheinlichkeit, dass ein Test positiv ausfällt, unter der Bedingung, dass eine tatsächliche

Krankheit vorliegt. Im Kontext von *Klassifizierungsverfahren* bezeichnet man die Sensitivität auch als *Recall*. Aus diagnostischer Sicht ist eine hohe Sensitivität wichtig, da man tatsächlich erkrankte Personen auch als krank diagnostizieren möchte.

Definition 6 (Spezifität)**.**
Sei $n_{(T=0|K=0)}$ die Anzahl der Personen mit einem negativen Testergebnis T und keiner tatsächlich vorliegenden Krankheit K. Des weiteren bezeichnet $n_{(K=0)}$ die Anzahl der Personen mit keiner tatsächlich vorhanden Krankheit K. So bezeichnen wir mit

$$P(T = 0|K = 0) = \frac{n_{(T=0|K=0)}}{n_{(K=0)}} = \frac{d}{b+d}$$

die (geschätzte) Spezifität eines Tests.

Somit schätzen wir die bedingte Wahrscheinlichkeit, dass ein Test negativ ausfällt, unter der Bedingung, dass keine tatsächliche Krankheit vorliegt. Hier sehen wir auch, dass die *Spezifität* eines Tests nichts Anderes als die *True-Negative-Fraction* ist. In diesem Kontext wird oft der Zusammenhang verwendet, dass $FNF = 1 - TNF$ und somit $FNF = 1 - \textit{Spezifität}$. Eine hohe Spezifität, also ein Test, welcher bei tatsächlich nicht erkrankten Individuen ein negatives Ergebnis liefert, ist ebenfalls gewünscht, da man Fehldiagnosen und die darauffolgenden Behandlungen vermeiden möchte.

Bei allgemeinen Klassifizierungsverfahren ist generell ein Trade-off zwischen der Spezifität und der Sensitivität zu beobachten. Mit dieser Frage beschäftigen wir uns nochmal kurz im Abschnitt der ROC-Kurven.

Bei diagnostischen Testverfahren wird eine stetige Variable y beispielsweise in Form eines Biomarkers erhoben. Dabei wird ein fester *Schwellenwert* t für die Klassifikation von einzelnen Individuen als krank oder gesund festgelegt. Dies bedeutet im Allgemeinen, dass wenn für ein Individuum der Wert y größer als t ist, dieses als krank klassifiziert wird. Führt man das Verfahren für alle N Individuen durch, so erhält man einen Datensatz, bei dem man die oben definierten *Fractions* bestimmen kann, gegeben für jedes Individuum liegen die Informationen über das Vorhandensein der tatsächlichen Krankheit vor. Dies werden wir im Unterabschnitt 4.2.1 detailliert am Beispieldatensatz *"Pancreatic cancer serum biomarkers study"* von Wieand et al. (1989) ausführen.

3. Regressionsmodelle

In diesem Kapitel möchten wir das Konzept der *intervall-zensierten* Daten einführen, besprechen geeignete Regressionsmodelle und wollen diese später mit den Ergebnissen aus diagnostischen Studien, also mit vorliegenden 4-Felder-Tafeln, in Verbindungen bringen. Zunächst werden wir ein paar grundlegende Resultate aus Tutz and Schmid (2016) behandeln.

3.1 Intervall-zensierte Daten und diskrete Funktionen

Im Allgemeinen liegen *intervall-zensierte* Daten vor, wenn man den Zeitpunkt des Auftretens eines Events nicht genau bestimmen kann, sondern lediglich das Intervall, im welchen das Event aufgetreten ist, angegeben werden kann. Dies führt uns zu gruppierten und somit diskreten Zeitpunkten an denen Events eingetreten sind. Das heißt, wir zerlegen unsere stetige Zeitachse in diskrete disjunkte Teilintervalle und betrachten dabei, ob ein Event im gegebenen Intervall aufgetreten ist oder nicht.

Definition 7 (Intervallzensierung, diskrete Zeitvariable)**.**
Sei nun

$$[0, a_1], (a_1, a_2], (a_2, a_3], \ldots, (a_{k-1}, a_k]$$

eine Zerlegung der stetigen Zeit $\mathcal{T}$ in disjunkte Intervalle. Weiterhin modelliere T die diskrete Zeitvariable, wobei Realisierung $T = t$ bezeichne, dass unser interessierendes Event im Zeitintervall $(a_{t-1}, a_t] \subset \mathbb{R}$ aufgetreten ist und $T \in \{1, \ldots, k\}$.

Man spricht hierbei auch von *grouped survival data* oder *discrete Time-to-Event data.*

Definition 8 (diskrete Hazardfunktion)**.**
Sei T eine diskrete Zeitvariable und $x^\top$ ein Vektor mit p erklärenden Variablen. Dann bezeichnen wir mit

$$\lambda(t|x) = P(T = t|T \geq t, x)$$

die diskrete Hazard- oder Risikofunktion für $t = 1, \ldots, k$. (engl. hazard function)

Bei der diskreten Hazardfunktion handelt es sich um die bedingte Wahrscheinlichkeit, dass ein Event im Intervall $(a_{t-1}, a_t]$ eintreten wird, gegeben das Intervall wurde erreicht. Somit beschreibt die Hazardfunktion die *unmittelbare Rate* für das Eintreten eines Events zum diskreten Zeitpunkt t, gegeben das Individuum hat bis dahin kein Event erfahren.

Dies führt uns zur zweiten wichtigen Funktion, welche mit der Hazardfunktion eng verbunden ist.

Definition 9 (diskrete Überlebenszeitfunktion)**.**
Sei T die diskrete Zeitvariable und $x^\top$ ein Vektor mit p erklärenden Variablen, dann bezeichnen wir mit

$$S(t|x) = P(T > t|x) = \prod_{i=1}^{t} (1 - \lambda(i|x))$$

die diskrete Überlebens(zeit-)funktion für $t = 1, \ldots, k$. (engl. survival function).

Die *survival function* bezeichnet somit die Wahrscheinlichkeit für das Eintreten eines Events nach dem Zeitpunkt t, beziehungsweise die Überlebenswahrscheinlichkeit im Intervall $(a_{t-1}, a_t]$. Wie man an der Definition sehen kann, multiplizieren wir die Gegenwahrscheinlichkeiten der Hazardfunktion zum jeweiligen Zeitpunkt $i = 1, \ldots, t$. Mit anderen Worten berechnen wir hiermit die Überlebenswahrscheinlichkeit als Wahrscheinlichkeit, kein Event im ersten Intervall zu erfahren, multipliziert mit der Wahrscheinlichkeit kein Event im zweiten Intervall zu erfahren und dies führt man bis zum Intervall t weiterhin fort.

3.2 Modellgleichung

Nach dem Buch von Tutz and Schmid (2016, Kapitel 3) interessieren wir uns nun dafür, die *hazard function* $\lambda(t|x)$ zum gegebenen Kovariablenvektor und bestimmten Zeitpunkten t zu parametrisieren, in dem man ein binäres Modell anpasst, welches zwischen dem Eintreten eines Events in Kategorie $\{t\}$ oder Kategorien $\{t+1, \ldots, k\}$, gegeben $T \geq t$, unterscheidet. Somit erhält man das diskrete Hazard-Modell

$$\lambda(t|x) = h(\gamma_{0_t} + x^\top \beta),$$

wobei $h(\cdot)$ eine *response function* bezeichnet, welche strikt monoton wachsend ist und die Parameter β die Einflüsse der jeweiligen Kovariablen darstellen. Somit stellen wir den Link zwischen dem Effekt unserer Kovariablen und unserer bedingten Wahrscheinlichkeit $\lambda(t|x)$ her. Man beachte hier, dass der Intercept γ_{0_t} nun von der Zeit abhängig ist und wir somit für jedes Intervall, welches durch t induziert wird, einen eigenen Parameter schätzen werden. Mittels der geforderten Eigenschaft der strikten Monotonie können wir die inverse Funktion $g = h^{-1}$ bestimmen, welche wir als *link function* bezeichnen werden. Das Modell hat dann die Form

$$g(\lambda(t|x)) = \gamma_{0_t} + x^\top \beta.$$

Verwenden wir für die *response function* die logistische Verteilungsfunktion $h(\eta) = \frac{exp(\eta)}{1+exp(\eta)}$, dann erhalten wir ein logistisches Regressionsmodell. Verwendet man stattdessen die Verteilungsfunktion der *Gompertz* Verteilung $h(\eta) = 1 - exp(-exp(\eta))$, so erhält man das Modell

$$\lambda(t|x) = 1 - exp(-exp(\gamma_{0_t} + x^\top \beta)).$$

Formt man dieses Modell noch ein wenig um

$$\begin{aligned} & 1 - \lambda(t|x) = exp(-exp(\gamma_{0_t} + x^\top \beta)) \\ \Leftrightarrow\ & log(-log(1 - \lambda(t|x))) = \gamma_{0_t} + x^\top \beta \\ \Leftrightarrow\ & log(-log(P(T > t | T \geq t, x))) = \gamma_{0_t} + x^\top \beta \end{aligned}$$

so erkennt man, dass es sich hier um eine *complementary log-log Link Function* handelt. Man spricht daher auch von einem *complementary log-log Modell* oder kurz geschrieben von einem *c-log-log Modell*. Es kann hierbei auch von einem *Grouped Proportional Hazards Model* gesprochen werden, da man es als diskrete Variante des Cox Proportional Hazard Modell betrachten kann, siehe hierfür Tutz and Schmid (2016, Kapitel 3.6). Nach Allison (1982) entsprechen die geschätzten β Parameter den Parametern aus dem Proportional Hazard Modell. Für die konkreten Schätzgleichungen der Parameter verweise ich ebenfalls auf Tutz and Schmid (2016, Kapitel 3.4).

3.3 Datengrundlage für das Regressionsmodell

Im Rahmen unseres Regressionsmodells müssen die Daten in einer entsprechenden Form bereitgestellt werden. Diese möchten wir im Folgenden genauer definieren.

Zunächst führen wir die binäre Variable

$$y_{is} = \begin{cases} 1 & \text{falls Individuum } i \text{ Event erlebt in } (a_{s-1}, a_s] \\ 0 & \text{falls Individuum } i \text{ überlebt in } (a_{s-1}, a_s] \end{cases}$$

für $s = 1, \ldots, t_i$ ein. Somit würde sich je Individuum i ein Beobachtungsvektor y_i mit $(0, \ldots, 0, 1)$ der Länge t_i ergeben. Das heißt, dass das Individuum die ersten $t_i - 1$ Intervalle überlebte und im Intervall t_i ein Event erfahren hat.

Betrachten wir erneut unseren linearen Prädiktor $\eta_{it} = \gamma_{0t} + x_i^\top \beta$ mit $t \in \{1, ..., k\}$, welchen wir auch wie folgt schreiben können

$$\eta_{it} = x_{it}^\top \tilde{\beta}$$

mit $\tilde{\beta}^\top = (\gamma_{01}, ..., \gamma_{0k}, \beta^\top)$. Nun können wir unsere *Design-Matrix* X_i wie folgt aufstellen:

$$y_i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad X_i \cdot \tilde{\beta} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & x_i^\top \\ 0 & 1 & 0 & \ldots & 0 & x_i^\top \\ 0 & 0 & 1 & \ldots & 0 & x_i^\top \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & x_i^\top \end{bmatrix} \cdot \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \\ \vdots \\ \gamma_{0k} \\ \beta \end{bmatrix}.$$

Eine Zeile in X_i beschreibt somit, in welchem Intervall sich das Individuum i befindet. Die Spaltenlänge hängt von der Länge des beobachteten Vektors y_i ab. Mit einer kleinen Modellierungsüberlegung lässt sich die *Design-Matrix* auch mittels einer Faktorvariable *Intervall* t darstellen:

$$X_i = \begin{bmatrix} \text{Intervall } 1 & x_i^\top \\ \text{Intervall } 2 & x_i^\top \\ \vdots & \vdots \\ \text{Intervall } t_i & x_i^\top \end{bmatrix}.$$

Somit können wir unseren Kovariablenvektor $x^\top$ um die Faktorvariable *Intervall* erweitern und bezeichnen den neuen Vektor mit $\tilde{x}^\top = (Intervall, x_1, \ldots, x_p)$. In unserer Anwendung werden wir sehen, dass wir mittels der *Software R* für jede Faktorausprägung der Variable *Intervall* einen eigenen Parameter schätzen und somit für jedes Zeitintervall einen spezifischen Intercept γ_{0t} erhalten werden.

3.4 Anwendung auf diagnostische Studien

Erinnern wir uns hier noch einmal an die Grundlagen von diagnostischen Studien und betrachten wir folgendes Problem. Bei diagnostischen Studien werden in der Praxis häufig mehrere verschiedene Schwellenwerte eines Biomarkers evaluiert und die Testergebnisse werden in Form von 4-Felder-Tafeln berichtet, jedoch auf Individualebene erhoben. Dies bedeutet, dass für jeden Studienteilnehmer der gemessene diagnostische Testwert vorliegt, aber in der Praxis meist nur Ergebnisse in aggregierter Form berichtet werden. Somit kann man die *True-Positive-Fraction* (Sensitivität) oder die *True-Negative-Fraction* (Spezifität)

schätzen. Möchte man jedoch eine Meta-Analyse durchführen, so liegt das Problem vor allem darin, dass man keine Individualdaten zur Verfügung hat und man sich lediglich auf die berichteten Schwellenwerte und die zugehörigen 4-Felder-Tafeln beziehen kann. Hilfreich sind deshalb Verfahren, die auch anhand von aggregierten Daten ROC-Kurven schätzen können und sich auf Meta-Analysen verallgemeinern lassen.

3.4.1 4-Felder-Tafeln und intervall-zensierte Daten

Gehen wir nun davon aus, dass uns keine Individualdaten mehr zur Verfügung stehen und wir lediglich die aggregierten Daten, also Anzahl an TP, TN, FP und FN sowie die Angabe des Schwellenwertes, vorliegen haben. Betrachten wir dazu die verschiedenen Schwellenwerte, auch Treshholds genannt, als unsere Zeitvariable t und erzeugen uns nach Hoyer et al. (2017) aus den 4-Felder-Tafeln *intervall-zensierte* Daten. Dazu betrachten wir im folgenden zwei Kennzahlen, zum einen die Anzahl an False Negatives sowie die Anzahl an True Negatives, oder mit anderen Worten, das Ereignis "negativ klassifiziert zu werden".

Zur Veranschaulichung nehmen wir die publizierten Daten von Choi et al. (2011). Bei dieser Studie handelt es sich um die Diagnose von Typ-2-Diabetes unter Verwendung des HbA1c-Wertes, welcher anhand einer Blutprobe bestimmt wird. Goldstandard ist der orale Glukosetoleranztest (OGTT). Die Daten liegen in Form von 4-Felder-Tafeln vor. Diese Tafeln werden in Abhängigkeit eines bestimmten Thresholds generiert. Auf die Festlegung der Thresholdwerte haben wir keinen Einfluss und können lediglich die bereitgestellten Kennzahlen mit Ihrem zugehörigen Thresholdwerten weiterverwenden. Auszugsweise schauen wir uns drei Tafeln mit den Thresholdwerten $5.0, 5.1$ und 5.2 an. Insgesamt berichtet die Studie 13 verschiedene HbA1c-Schwellenwert und damit 13 Paare von Sensitivität und Spezifität.

Threshold $t = 5.0$:

	Krank	Gesund	
Positive	617	7735	
Negative	18	1005	
	635	8740	9375

Threshold $t = 5.1$:

	Krank	Gesund	
Positive	607	7123	
Negative	28	1617	
	635	8740	9375

Threshold $t = 5.2$:

	Krank	Gesund	
Positive	600	6302	
Negative	35	2438	
	635	8740	9375

Behandeln wir nun unsere Thresholdvariable als unsere Zeitvariable t. Betrachten wir im Folgenden die gegeben Thresholdwerte als obere Intervallgrenze, so können wir unserer Thresholdsvariable in Intervalle

$$(-\infty, 0.0], (0.0, 5.0], (5.0, 5.1], (5.1, 5.2], \ldots, (6.6, \infty)$$

zerlegen. Das erste sowie das letzte Intervall sind für unser späteres Modell aus rein technischer Natur aufgeführt, da wir die Kennzahlen künstlich hinzufügen können. Wir wissen beispielsweise zu einem Schwellenwert von $t = 0.0$, dass $TP = 635$ und $FP = 8740$ sein muss, analog gilt für $t = \infty$ $TN = 8740$ sowie $FN = 635$.

3.4.2 Intervallzensierung und Events

Beschränken wir uns zunächst nur auf die Individuen der Gruppe *Krank*, so können wir in der 4-Felder-Tafel mit $t = 5.0$ die Werte $TP = 617$ und $FN = 18$ ablesen. Betrachten wir dies nun im Zusammenhang mit den oben generierten Intervallen, so können wir feststellen, dass zum Zeitpunkt $t = 5.0$ 18 Individuen vorliegen, welche fälschlicherweise als negativ klassifiziert wurden. Das heißt, dass diese Personen einen diagnostischen Testwert innerhalb des Intervalls $I_1 = (0, 5.0]$ haben. Mit anderen Worten haben diese Individuen ein Event innerhalb des Intervalls $(0, 5.0]$ erlebt, nämlich "als negativ klassifiziert zu werden". Da unsere Thresholdvariable nur steigen kann, so sind diese Individuen auch für alle weiteren Intervalle $(5.0, 5.1], (5.1, 5.2], \ldots, (6.6, \infty)$ als negativ klassifiziert. Somit können wir die Kennzahl $FN = 18$ als die Gesamtanzahl der Events bis zum betrachteten Intervall verstehen. Die Kennzahl $TP = 617$ hingegen bezeichnet die Individuen *unter Risiko*, also die Anzahl an Individuen, welche noch die Möglichkeit haben ein Event in einem der kommenden Intervalle zu erleben.

Betrachtet man nun die 4-Felder-Tafel zu $t = 5.1$, also das Intervall $I_2 = (5.0, 5.1]$, so haben wir $TP = 607$ und $FN = 28$, also bis zu $t = 5.1$ erhalten wir 28 falsch negativ klassifizierte Individuen. Ziehen wir die Individuen, die bereits ein Event im vorherigen Intervall hatten ab, so haben wir $28 - 18 = 10$ neue Events im Intervall I_2. Wie oben haben wir $TP = 617 - 10 = 607$ Individuen *unter Risiko*.

Zusammenfassend können wir den Zuwachs der FN zwischen zwei aufeinanderfolgenden Intervallen als Anzahl der Events im Sinne der Ereigniszeitanalyse auffassen. *Überleben* bedeutet dann, kein Event in diesem Intervall sowie in allen vorherigen Intervallen erlebt zu haben. Dies ist gleichbedeutend mit unserer *Sensitivität*, also dem Anteil der TP an der Gesamtzahl der Individuen in der Gruppe *Krank* zum entsprechenden Thresholdwert t.

Betrachten wir nun auch die Individuen der Gruppe *Gesund*, so können wir mit der gleichen Argumentation wie zuvor das Ereignis "negativ klassifiziert zu werden" als Event handhaben. Für $t = 5.0$ haben wir $TN = 1005$, diese Kennzahl entspricht der Anzahl richtig negativ klassifizierter Individuen, das heißt diese können mit steigendem Threshold kein weiteres Event erleben. Dabei stellen die $FP = 7735$ wieder die Individuen *unter Risiko* dar, also diejenigen die mit steigendem Threshold noch ein Event erfahren könnten. In diesem Fall bedeutet *Überleben* weiterhin als False Positive klassifiziert zu werden. Dies entspricht unserem Schätzer für die *False-Positive-Fraction*, welcher aus dem Verhältnis von FP und Anzahl der *Gesunden* bestimmt wird. Daraus können wir die *Spezifität* als $1 - FPF$ ableiten.

Schließlich haben wir gesehen, dass man bei entsprechender Datenaufbereitung die vorhandenen 4-Felder-Tafeln als *intervall-zensierte* Daten auffassen kann. Dies ermöglicht uns das bereits eingeführte *complementary-log-log Modell* zu verwenden und darüber einen Schätzer für die diskrete Hazardfunktion zum jeweiligen Threshold t zu ermitteln und über Definition 9 auch einen Schätzer für die *Sensitvität* sowie für die *Spezifität* abzuleiten. Dies werden wir im weiteren Verlauf dieser Arbeit noch genauer besprechen.

3.5 Datenaufbereitung

Im weiteren Verlauf werden wir folgende Notationen verwenden. Die Anzahl der Individuen aus der Gruppe *Krank* bezeichnen wir mit *Kranke* und die Individuen aus der Gruppe *Gesund* entsprechend mit *Gesunde*.

Wie im Unterabschnitt 3.4.1 gezeigt wurde, können wir uns aus den 4-Felder-Tafeln gruppierte *intervall-zensierte* Daten erzeugen. Das heißt, wir können für jedes Intervall $(a_i, a_{i+1}]$ die Anzahl der Events bei den *Kranken* sowie den *Gesunden* ermitteln. Nun möchten wir mit einer entsprechenden Datenaufbereitung das *complementary-log-log Modell* für die Schätzung der Hazardrate verwenden. Dafür müssen wir die gruppierten Daten in Pseudobeobachtungen auflösen. Das heißt, dass wir für jedes Event ein Pseudoindividuum generieren, welches bis zum gegebenen Threshold kein Event erlebt hat.

Betrachten wir nun Tabelle 3.5.1. Die Daten entsprechen wie zuvor den Daten aus der Studie von Choi et al. (2011). Eine Zeile entspricht einer 4-Felder-Tafel. Zusätzlich wurde die Anzahl neuer Events zum Intervall t wie folgt hinzugefügt. Wie bereits festgestellt wurde, entsprechen die dazugekommenen FN den Events bei den *Kranken* und die dazugekommenen TN den Events bei den *Gesunden*. Somit bezeichnen die Spalten *FN_Evt* und *TN_Evt* die jeweilige Anzahl der neu aufgetretenen Events im Intervall t. Die Spalten *FN_NoEvt* und *TN_NoEvt* bezeichnen die Anzahl der kranken sowie gesunden Individuen, welche kein Event bis zum Threshold t erfahren haben.

TP	TN	K	G	t	FN	FP	FN_Evt	TN_Evt	FN_NoEvt	TN_NoEvt
635	0	635	8740	0.0	0	8740	0	0	635	8740
617	1005	635	8740	5.0	18	7735	18	1005	617	7735
607	1617	635	8740	5.1	28	7123	10	612	607	7123
600	2438	635	8740	5.2	35	6302	7	821	600	6302
581	3409	635	8740	5.3	54	5331	19	971	581	5331
563	4422	635	8740	5.4	72	4318	18	1013	563	4318
550	5384	635	8740	5.5	85	3356	13	962	550	3356
522	6267	635	8740	5.6	113	2473	28	883	522	2473
489	6966	635	8740	5.7	146	1774	33	699	489	1774
457	7534	635	8740	5.8	178	1206	32	568	457	1206
429	7927	635	8740	5.9	206	813	28	393	429	813
393	8172	635	8740	6.0	242	568	36	245	393	568
332	8460	635	8740	6.2	303	280	61	288	332	280
236	8670	635	8740	6.6	399	70	96	210	236	70
0	8740	635	8740	∞	635	0	236	70	0	0

Tabelle 3.5.1: Gruppierte Events zu Threshold t

Beispielsweise haben bis zum Threshold $t = 5.4$ insgesamt 72 Individuen ein Event erlebt, jedoch waren es nur 18 Individuen innerhalb des Intervalls $I_{t=5.4} = (5.3, 5.4]$.

Nun betrachten wir Tabelle 3.5.2, um noch eine weitere wichtige Kennzahl zu ermitteln. Zu dem jedem Threshold t bleiben natürlicherweise die Anzahlen der *Kranken* K sowie *Gesunden* G gleich, daher können wir diese zwei Spalten vorerst ignorieren und behalten die Zahlen im Hinterkopf. Nun müssen wir uns zwei Spalten hinzufügen, welche die Individuen *unter Risiko* in der Gruppe der *Kranken* bzw. *Gesunden* bereitstellen. Die Anzahl *unter Risiko* entspricht genau der Anzahl der Individuen, welche im vorherigen Threshold t noch kein Event erfahren haben. Somit haben wir in der neuen Tabelle die Spalten

K_Risk, welche die Anzahl der kranken Individuen, die kein Event im vorherigen Intervall erlebten (also nicht FN sind), sowie *G_Risk*, welche die Anzahl der gesunden Individuen, die kein Event im vorherigen Intervall erlebten (also nicht TN sind), hinzugefügt.

TP	TN	t	FN	FP	FN_Evt	TN_Evt	FN_NoEvt	TN_NoEvt	K_Risk	G_Risk
635	0	0.0	0	8740	0	0	635	8740	635	8740
617	1005	5.0	18	7735	18	1005	617	7735	635	8740
607	1617	5.1	28	7123	10	612	607	7123	617	7735
600	2438	5.2	35	6302	7	821	600	6302	607	7123
581	3409	5.3	54	5331	19	971	581	5331	600	6302
563	4422	5.4	72	4318	18	1013	563	4318	581	5331
550	5384	5.5	85	3356	13	962	550	3356	563	4318
522	6267	5.6	113	2473	28	883	522	2473	550	3356
489	6966	5.7	146	1774	33	699	489	1774	522	2473
457	7534	5.8	178	1206	32	568	457	1206	489	1774
429	7927	5.9	206	813	28	393	429	813	457	1206
393	8172	6.0	242	568	36	245	393	568	429	813
332	8460	6.2	303	280	61	288	332	280	393	568
236	8670	6.6	399	70	96	210	236	70	332	280
0	8740	∞	635	0	236	70	0	0	236	70

Tabelle 3.5.2: Gruppierte Events, Individuen unter Risiko

Als Nächstes möchten wir *intervall-zensierte* Daten generieren und dafür müssen wir für jedes Intervall $(-\infty, 0], (0, 5.0], (5.0, 5.1], (5.1, 5.2], \ldots, (6.6, \infty)$ Pseudobeachtungen einführen. Insgesamt haben wir $N = 9375$ Individuen und für jedes Individuum müssen wir einen Pseudovektor generieren. Bereits in Abschnitt 3.2 haben wir gesehen, dass jedem Individuum i ein Beobachtungsvektor $y_i = (0, \ldots, 0, 1)$ zugeordnet wird. Somit müssen wir für jedes der oben genannten Intervalle feststellen, ob das Individuum in diesem Intervall ein Event erfahren hat oder nicht und aus diesen Informationen erstellen wir für jedes Individuum einen entsprechenden Pseudobeobachtungsvektor y_i.

Nun kennen wir für jedes Intervall die Zuwächse an Events, also die FN_Evt und TN_Evt. So müssen wir entsprechend die 9375 Individuen mit den passenden Pseudovektoren versehen und erhalten somit einen *intervall-zensierten* Datensatz, bei dem wir für jedes Individuum über den Beobachtungsvektor bestimmen können, in welchem Intervall das Event eingetreten ist. Für die genauere Datenaufbereitung verweise ich auf meinen erstellten R-Code zur Generierung von Pseudobeobachtungen.

Beispielsweise haben wir in Tabelle 3.5.3 die Datenmatrix für das Individuum $i = 35$

ID	outcome	intervall	group	intcode
35	0	$(-\infty, 0.0]$	0	1
35	0	$(0.0, 5.0]$	0	2
35	0	$(5.0, 5.1]$	0	3
35	1	$(5.1, 5.2]$	0	4

Tabelle 3.5.3: Datenmatrix vom Individuum i=35

und Tabelle 3.5.4 stellt entsprechend die Datenmatrix für das Individuum $i = 6350$ dar.

Hierbei bezeichnet die Spalte *outcome*, ob ein Event eingetreten ist oder nicht. Die Variable *intcode* kodiert das entsprechende Intervall, so steht für *intcode* $= 2$ das Intervall $I_2 =$

	ID	outcome	intervall	group	intcode
1	6350	0	$(-\infty, 0.0]$	1	1
2	6350	0	$(0.0, 5.0]$	1	2
3	6350	0	$(5.0, 5.1]$	1	3
4	6350	0	$(5.1, 5.2]$	1	4
5	6350	0	$(5.2, 5.3]$	1	5
6	6350	0	$(5.3, 5.4]$	1	6
7	6350	0	$(5.4, 5.5]$	1	7
8	6350	1	$(5.5, 5.6]$	1	8

Tabelle 3.5.4: Datenmatrix vom Individuum i=6350

$(0, 5.0]$. Dies ist, wie in Abschnitt 3.2 beschrieben, wichtig, da wir mit der Faktorvariable *intcode* unseren spezifischen Intercept γ_{0r} schätzen werden. Die Spalte *group* gibt uns an, ob der Patient zu den *Gesunden* $\widehat{=}$ 1 oder zu den *Kranken* $\widehat{=}$ 0 gehört. Dies ist später für unser Modell notwendig, um einen Schätzer für die *Senstivität* bzw. *Spezifität* zu erhalten.

3.6 Modellanwendung

Verwenden wir nun unser *complementray log-log Modell* aus Abschnitt 3.2

$$\begin{aligned} & \lambda(t|x) = 1 - exp(-exp(\gamma_{0_t} + x^\top \beta)) \\ \Leftrightarrow \; & \lambda(t|x) = 1 - exp(-exp(\tilde{x}^\top \tilde{\beta})), \end{aligned}$$

wobei γ_{0_t} unseren intervallspezifischen Parameter, $\tilde{\beta}^\top = (\gamma_{01}, ..., \gamma_{0k}, \beta^\top)$ unseren erweiterten Parametervektor und $\tilde{x}^\top = (Intervall, x_1, \ldots, x_p)$ unseren erweiterten Kovariablenvektor darstellen.

In dieser Anwendung haben wir nach unseren Überlegungen den Kovariablenvektor $\tilde{x}^\top = (intcode, group)$ eingeführt. Die Faktorvariable *incode* erlaubt es uns, einen Parameter für jedes Intervall zu schätzen, und die binär-kodierte Variable *group* gibt uns die Gruppenzugehörigkeit *Krank* oder *Gesund* wieder. Hierbei sei noch anzumerken, dass bei der computationalen Berechnung im Hintergrund jede einzelne Ausprägung beider Variablen als dummy-kodierte Einflussgröße aufgefasst wird.

3.6.1 Schätzung der Hazard- und Survivalfunktion

Mit dem oben eingeführten Modell und dem Kovariablenvektor $\tilde{x}^\top$ ist es uns möglich, die Hazardfunktion, also das Risiko für das Auftreten eines Events in einem bestimmten Intervall, zu schätzen

$$\hat{\lambda}(t|x) = 1 - exp(-exp(\tilde{x}^\top \hat{\tilde{\beta}})).$$

Dabei bezeichnet $\hat{\tilde{\beta}}^\top = (\hat{\gamma_{01}}, ..., \hat{\gamma_{0k}}, \hat{\beta}^\top)$, wobei die ersten k Parameter, also $\hat{\gamma_{01}}, ..., \hat{\gamma_{0k}}$, den intervallspezifischen Intercepts entsprechen. Diese werden in der *Software R* mittels einer Dummy-Kodierung über die eingeführte Faktorvariable geschätzt. Die restlichen Parameter in $\hat{\beta}$ setzen sich aus dem Interaktionseffekt $\hat{\beta}_{intcode:group}$ zusammen. Letzteres ist insoweit sinnvoll, da wir zwischen den zwei Gruppen einen unterschiedlichen, also gruppenspezifischen, Anstieg an Events erwarten.

Möchten wir nun die Überlebenszeitfunktion, also die *survival function* für ein bestimmtes Intervall schätzen, so nutzen wir die Definition 9 und erhalten

$$\hat{S}(t|\tilde{x}) = \prod_{i=1}^{t}(1 - \hat{\lambda}(i|\tilde{x})) = \prod_{i=1}^{t} exp(-exp(\tilde{x}^\top \hat{\tilde{\beta}})).$$

Als Nächstes betrachten wir die Schätzer für die jeweiligen Gruppen, also für *Kranke* und *Gesunde*. Da wir unsere Kovariable *group* binär kodiert haben, werden zur Schätzung der *survival function* die gruppenspezifischen Parameter $\beta_{group:intcod}$ für die Gruppe der *Kranken* durch die Ausprägung der binären Variable mit 0 multipliziert. Somit ergibt sich beispielsweise für folgende Kovariablenausprägung $\tilde{x}^\top = (intcode = 1, group = 0)$ die Gleichung

$$\begin{aligned}\hat{\lambda}(t|(1,0)^\top) &= 1 - exp(-exp(1 \cdot \hat{\gamma_{01}} + 1 \cdot 0 \cdot \hat{\beta}_{1,intcod:group)})) \\ &= 1 - exp(-exp(\hat{\gamma_{01}}))\end{aligned}$$

und für $\tilde{x}^\top = (intcode = 2, group = 0)$

$$\hat{\lambda}(t|(2,0)^\top) = 1 - exp(-exp(\hat{\gamma_{02}})).$$

Betrachten wir nun die den Schätzer für die Ausprägung $\tilde{x}^\top = (intcode = 1, group = 1)$ und erhalten somit

$$\hat{\lambda}(t|(1,1)^\top) = 1 - exp(-exp(\hat{\gamma_{01}} + \hat{\beta}_{1,intcod:group}))$$

und ebenfalls für $\tilde{x}^\top = (intcode = 2, group = 1)$

$$\hat{\lambda}(t|(1,1)^\top) = 1 - exp(-exp(\hat{\gamma_{02}} + \hat{\beta}_{2,intcod:group})).$$

Möchten wir nun die *survival function* für die jeweiligen Gruppen im zweiten Intervall $\widehat{=} \{intcode = 2\}$ bestimmen, so setzen wir unsere geschätzten Hazards $\hat{\lambda}_t$ in die oben genannte Gleichung ein und erhalten für $group = 0$

$$\hat{S}(t|group) = exp(-exp(\hat{\gamma_{01}})) \cdot exp(-exp(\hat{\gamma_{02}}))$$

und äquivalent erhält man für $group = 1$

$$\begin{aligned}\hat{S}(t|group) = &exp(-exp(\hat{\gamma_{01}} + \hat{\beta}_{1,intcod:group})) \\ &\times exp(-exp(\hat{\gamma_{02}} + \hat{\beta}_{2,intcod:group})).\end{aligned}$$

3.6.2 Schätzung von Sensitivität und Spezifität

Erinnern wir uns an die Definition der *survival function*, also an folgende Wahrscheinlichkeit $S(t|x) = P(T > t|x)$. Sie bezeichnet somit die Wahrscheinlichkeit für das Eintreten eines Events nach dem Zeitpunkt t, beziehungsweise die Wahrscheinlichkeit im Intervall $I_t = (a_{t-1}, a_t]$ zu überleben. Wie wir bereits in Abschnitt 3.4 gesehen haben, ist die Überlebenswahrscheinlichkeit im Intervall I_t für die Gruppe der *Kranken*, also $group = 0$, gleichbedeutend mit unserer gesuchten *Sensitivität*, weil wir damit den Anteil der TP an der Gruppe der *Kranken* schätzen. Somit kann die Sensitivität wie folgt parametrisiert werden:

$$Sens(t) = TPF(t) = \prod_{i=1}^{t} exp(-exp(\hat{\gamma_{0i}}))$$

Analog verfährt man bei der Schätzung der *Spezifität*. Hierbei setzt man die Variable *group* $= 1$ und schätzt somit zuerst die Überlebenswahrscheinlichkeit, also die Wahrscheinlichkeit noch nicht negativ klassifiziert worden zu sein unter den *Gesunden*. Dies ist gleichbedeutend mit der Wahrscheinlichkeit False Positive klassifiziert worden zu sein. Dies führt uns dann zum Schätzer für die *False-Positive-Fraction*, also des Anteils der FP an den *Gesunden* und somit erhalten wir

$$FPF(t) = \prod_{i=1}^{t} exp(-exp(\hat{\gamma_{0i}} + \hat{\beta}_{i,intcod:group})).$$

Daraus können wir die *Spezifität* sofort herleiten durch

$$Spez(t) = 1 - FPF(t).$$

3.6.3 Auswertung

Für die Auswertung unseres Beispieldatensatzes aus Abschnitt 3.5 passen wir nun ein GLM mit *complementary log-log* Link an. Die folgenden Auswertungen können mittels der R-Datei `main_data_process.R` reproduziert werden. Hierbei ist zu beachten, dass wir den Intercept mit -1 entfernen, da unsere intervallspezifischen Intercepts bereits mit der Variable *intcod* in das Modell einfließen, vergleiche hierzu Abschnitt 3.2.

```
# GLM mit c-log-log Link
mod <- glm(outcome ~ -1 + intcod + group : intcod,
           family = binomial(link = "cloglog"), data = df)
summary(mod)
```

In Tabelle 3.6.1 können wir die Regressionsparameter aus unserem *summary output* herauslesen, hierbei sehen wir für *intervall 1* und *intervall 15* zwei eher technische Resultate, welche wir bereits in Abschnitt 3.4 erwähnt haben. Berechnen wir mit unseren geschätzten Parametern beispielhaft den Schätzer für die *Sensitivität*, also *group* $= 0$, zum Zeitpunkt $t = 5.1$ also für das dritte Intervall

$$\begin{aligned} Sens(5.1) &= exp(-exp(-18.476)) \cdot exp(-exp(-3.549)) \cdot exp(-exp(-4.114)) \\ &= 0.955906 \end{aligned}$$

Blicken wir nochmal in Abschnitt 3.4 zu unseren 4-Felder-Tafeln zurück. Wie man bei Threshold $t = 5.1$ ablesen kann, erhalten wir eine (deskriptive) *Sensitivität* von

$$\frac{\#\ True\ Positive}{\#\ Kranke} = \frac{607}{635} = 0.9559055.$$

Aus dem Modell berechnen wir somit genau die *Sensitivität*, welche wir auch deskriptiv aus den 4-Felder-Tafeln ermitteln können. Der Vorteil unserer Methode liegt aber darin, dass wir die Sensitivität in Abhängigkeit eines Regressionsmodells berechnen können und somit auch dessen Vorteile durch die Parametrisierung erhalten. Desweiteren ist es von Interesse das angeführte Modell im Kontext der Meta-Analyse verwenden zu können. Hierbei wäre es denkbar, das Modell zu einem gemischten Modell zu erweitern.

Analog wie zuvor führen wir noch eine beispeilhafte Berechnung für die *Spezifität* zum Zeitpunkt $t = 5.1$ durch

$$\begin{aligned} FPF(5.1) = exp(-exp(-18.476)) &\cdot exp(-exp(-3.549 + 1.446)) \\ &\times exp(-exp(-4.114 + 1.618)) = 0.8149885. \end{aligned}$$

	Dependent variable:
	outcome
intervall 1 $(-\infty, 0.0]$	-18.476 (247.474)
intervall 2 $(0.0, 5.0]$	-3.549^{***} (0.236)
intervall 3 $(5.0, 5.1]$	-4.114^{***} (0.316)
intervall 4 $(5.1, 5.2]$	-4.457^{***} (0.378)
intervall 5 $(5.2, 5.3]$	-3.436^{***} (0.229)
intervall 6 $(5.3, 5.4]$	-3.459^{***} (0.236)
intervall 7 $(5.4, 5.5]$	-3.757^{***} (0.277)
intervall 8 $(5.5, 5.6]$	-2.952^{***} (0.189)
intervall 9 $(5.6, 5.7]$	-2.729^{***} (0.174)
intervall 10 $(5.7, 5.8]$	-2.693^{***} (0.177)
intervall 11 $(5.8, 5.9]$	-2.761^{***} (0.189)
intervall 12 $(5.9, 6.0]$	-2.434^{***} (0.167)
intervall 13 $(6.0, 6.2]$	-1.780^{***} (0.128)
intervall 14 $(6.2, 6.0]$	-1.075^{***} (0.103)
intervall 15 $[6.6, \infty)$	2.990 (42.903)
intervall 1 $(-\infty, 0.0]$:group 1	0.000 (256.306)
intervall 2 $(0.0, 5.0]$:group 1	1.446 (0.238)
intervall 3 $(5.0, 5.1]$:group 1	1.618 (0.319)
intervall 4 $(5.1, 5.2]$:group 1	2.357 (0.380)
intervall 5 $(5.2, 5.3]$:group 1	1.649 (0.232)
intervall 6 $(5.3, 5.4]$:group 1	1.902 (0.238)
intervall 7 $(5.4, 5.5]$:group 1	2.379 (0.279)
intervall 8 $(5.5, 5.6]$:group 1	1.765 (0.192)
intervall 9 $(5.6, 5.7]$:group 1	1.627 (0.178)
intervall 10 $(5.7, 5.8]$:group 1	1.741 (0.182)
intervall 11 $(5.8, 5.9]$:group 1	1.830 (0.196)
intervall 12 $(5.9, 6.0]$:group 1	1.409 (0.179)
intervall 13 $(6.0, 6.2]$:group 1	1.434 (0.142)
intervall 14 $(6.2, 6.0]$:group 1	1.402 (0.127)
intervall 15 $[6.6, \infty)$:group 1	0.000 (89.700)
Observations	66,475
Akaike Inf. Crit.	46,201.920
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabelle 3.6.1: c-log-log Modellparameter

Und somit erhalten wir auch

$$Spez(5.1) = 1 - FPF(5.1) = 0.1850115.$$

Weiterhin sind hier nur gerundete Parameterwerte angegeben, die Berechnungen wurden jedoch mit ungerundeten Werten aus dem Modelloutput mittels *Software R* durchgeführt.

Verlgeichen wir dieses Ergebnis wieder mit unserer (deskriptiven) *Spezifität* aus der 4-Felder-Tafel zum Threshold $t = 5.1$

$$\frac{\#\ True\ Negative}{\#\ Gesunde} = \frac{1617}{8740} = 0.1850114,$$

so erhalten wir ebenfalls das gleiche Ergebnis.

3.7 Life-Tables und ROC-Kurve

In diesem Abschnitt möchten wir unsere Ergebnisse visualisieren, dazu bedienen wir uns weiterhin der Methoden aus der Ereigniszeitanalyse. Dafür werden wir Life-Tables und ROC-Kurven verwenden.

3.7.1 Life-Tables

Die sogenannten *Life-Tables* sind eine Veranschaulichung der Veränderung von Überlebenswahrscheinlichkeiten über die Zeit hinweg. In unserer Anwendung interessieren wir uns dafür, wie sich die *Sensitivität* bzw. die *Spezifität* über den Anstieg des Thresholds t verändern.

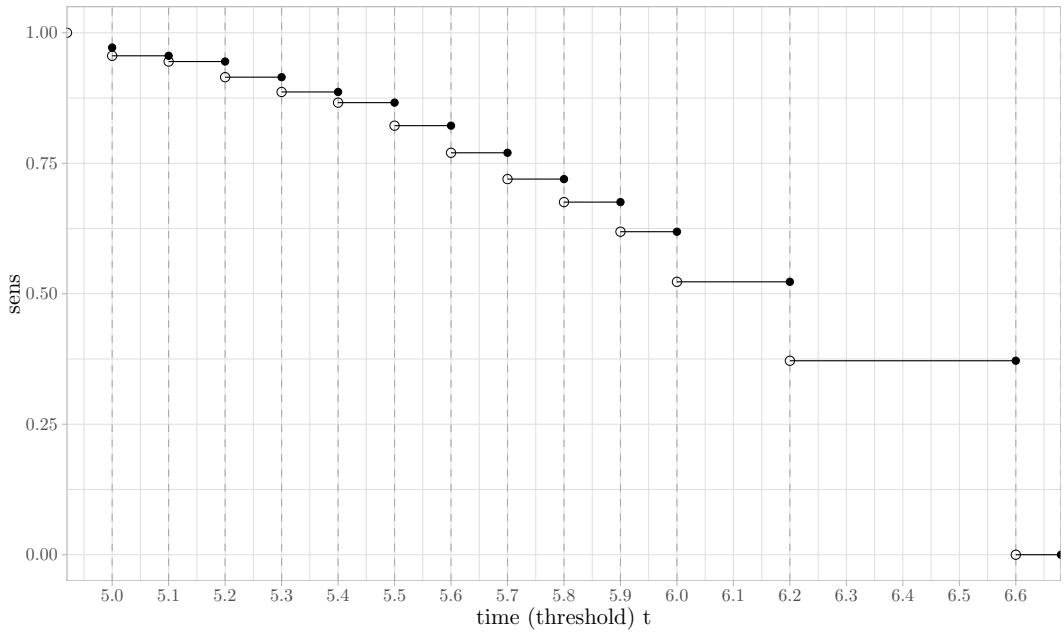

Abbildung 3.7.1: Life-Table - Sensitiviät aus c-log-log Modell

In Abbildung 3.7.1 sehen wir die *Sensivität* in Abhängigkeit der Thresholdvariable t. Die offenen und geschlossenen Kreise sowie die Verbindungslinie symbolisieren unsere halboffenen Intervalle der Form $(a_i, a_{i+1}]$. Somit wird die Intervallzensierung berücksichtigt. Durch diese Treppenfunktion können wir unter Berücksichtigung der Intervalle eine Schätzung

für einen gegebenen Thresholdwert ablesen. Wie wir sehen können, sinkt die *Sensitvität* mit steigendem t. Also sinkt die Wahrscheinlichkeit positiv klassifiziert zu werden unter Bedingung, dass man zu *Kranken* gehört.

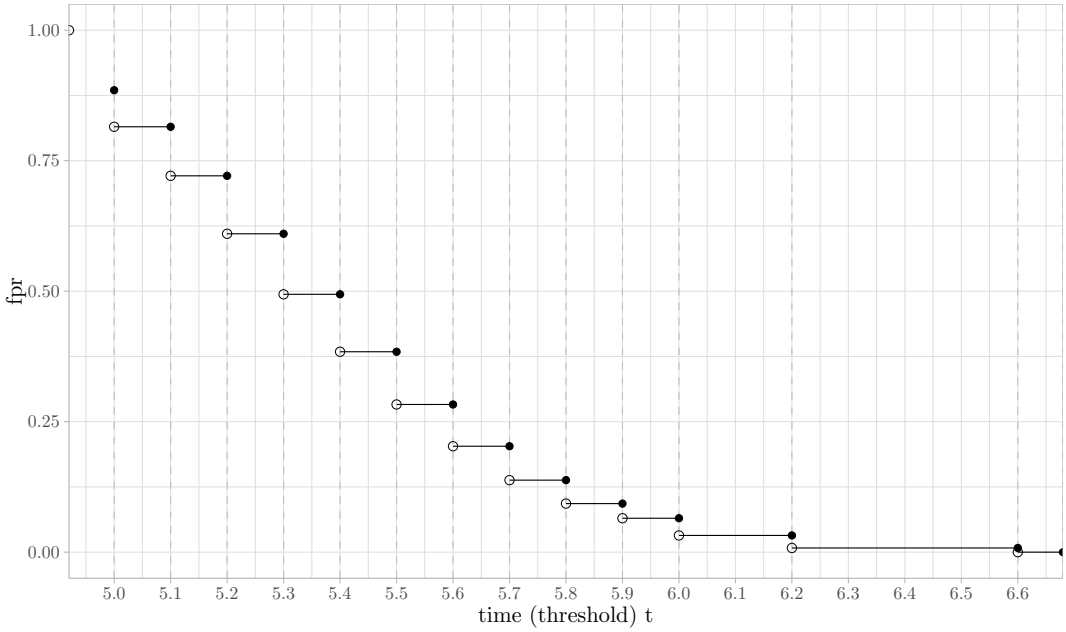

Abbildung 3.7.2: Life-Table - False-Positive-Fraction aus c-log-log Modell

In Abbildung 3.7.2 können wir die *False-Positive-Fraction* in Abhängigkeit unserer Thresholdvariable t ablesen. Diese entspricht, wie in den vorherigen Abschnitten besprochen, der Überlebenswahrscheinlichkeit zum Zeitpunkt t. Hierbei können wir erkennen, dass die *False-Positive-Fraction* schneller als die *Sensitivität* gegen die 0 verläuft. Es ist also bei niedrigeren Thresholdwerten wahrscheinlicher als positiv klassifiziert zu werden wenn man eigentlich zu *Gesunden* gehört, als wenn man zu *Kranken* gehört. Bei einem Thresholdwert von $t > 5.3$ unterschreitet die Überlebenswahrscheinlichkeit bei *Gesunden* den Wert von 0.5. Die Wahrscheinlichkeit als *False Positive* klassifiziert zu werden mit einem Thresholdwert von mindestens 5.4 liegt demnach bei unter 50%.

In Abbildung 3.7.3 können wir die *Spezifität* in Abhängigkeit unserer Thresholdvariable t ablesen. Hierbei haben wir die *Spezifität* über $Spez(t) = 1 - FPF(t)$ berechnet und abgetragen. Wir erkennen, dass die Wahrscheinlichkeit *True Negative* klassifiziert zu werden mit wachsendem Thresholdwert steigt.

3.7.2 ROC-Kurve

Abschließend zu den grafischen Visualisierungen möchten wir noch die wichtigste Darstellungsmöglichkeit, nämlich die *ROC-Kurve*, verwenden. Bei dieser Methode tragen wir die Schätzer für *Sensitivität* sowie für *1 - Spezifität* gegeneinander ab und verbinden diese zur Veranschaulichung mittels abschnittsweiser Linien.

In Abbildung 3.7.4 entspricht ein abgetragener Punkt einem Tupel $(Sens, 1 - Spez)$ zu einem vorgegebenen Thresholdwert t. In Tabelle 3.7.1 finden wir einen Auszug aus der gegeben Datenlage, wobei t unserem Thresholdwert, *sens* unserer aus dem Modell geschätzten *Sensitivität* und *spez* der ebenfalls aus dem Modell geschätzten *Spezifität* entspricht. Wie

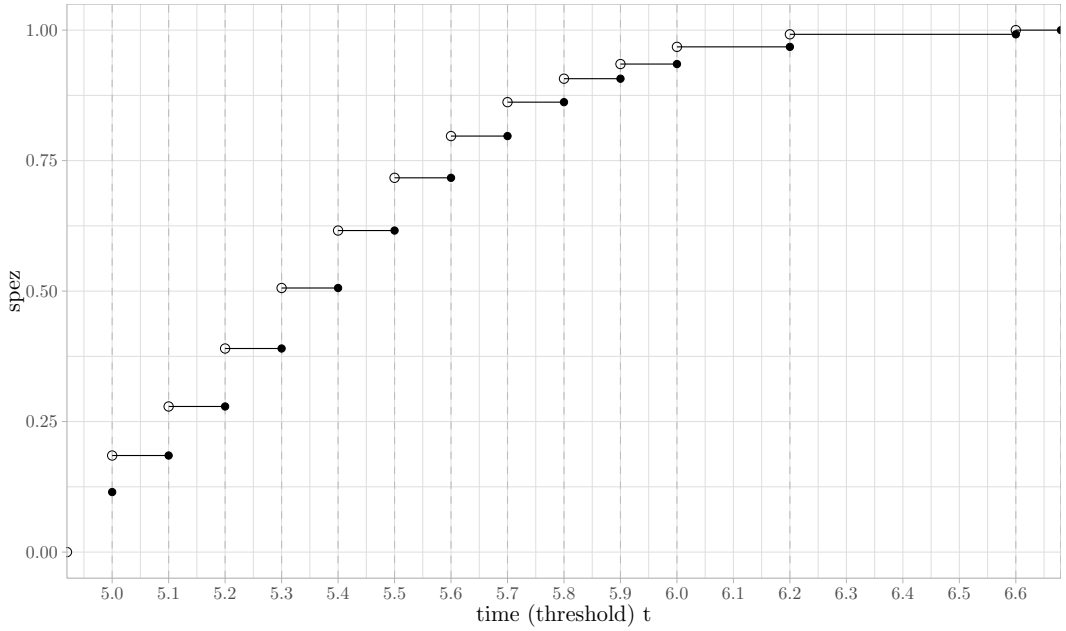

Abbildung 3.7.3: Spezifität aus c-log-log Modell

wir aus der Tabelle ablesen können, müssen wir bei der Interpretation der *ROC-Kurve* in Bezug auf unsere Thresholdwerte auf die richtige Richtung achten. Möchte man in Bezug auf aufsteigende Werte von t interpretieren, so startet man in Abbildung 3.7.4 von rechts nach links. Beispielsweise haben wir für $t = 6.2$, also im Intervall $(6.1, 6, 2]$, das Tupel $(0.37, 0.01)$, also eine Wahrscheinlichkeit von 0.37 als *True Positive* klassifiziert zu werden unter der Bedingung krank zu sein und eine 0.01 Wahrscheinlichkeit als *False Positive* klassifiziert zu werden, unter der Bedingung gesund zu sein.

t	sens	1 - spez
0.0	1.00	1.00
5.0	0.97	0.89
5.1	0.96	0.81
⋮	⋮	⋮
6.2	0.52	0.03
6.6	0.37	0.01
∞	0.00	0.00

Tabelle 3.7.1: Auszug zur ROC-Kurve

Generell zur Verwendbarkeit einer *ROC-Kurve* sollte man erwähnen, dass man je nach Kostenfaktor unterschiedliche Ziele verfogt. Für eine medizinische Diagnose oder für ein Screening-Programm möchte man eine sehr hohe *Sensitivtät* haben, da es bei einer schwerwiegenden Krankheit deutlich wichtiger ist einer Person ein positives Testergebnis geben zu können, wenn diese auch wirklich krank ist. Wobei man dann unter Umständen die hohe Anzahl an falsch positiven Testergebnissen in Kauf nehmen müsste. Hierbei kommt aber wieder der Kostenfaktor ins Spiel, beispielsweise kann nach einer falsch-positiven Diagnose eine sehr lange und risikoreiche Therapie erfolgen, so ist dies ebenfalls ein Faktor, den man hierbei mitberücksichtigen muss. Eine Möglichkeit um die Spezifität hoch zu halten, also die *False-Positive-Fraction* zu minimieren, ist es, im Voraus schon eine Selektierung der

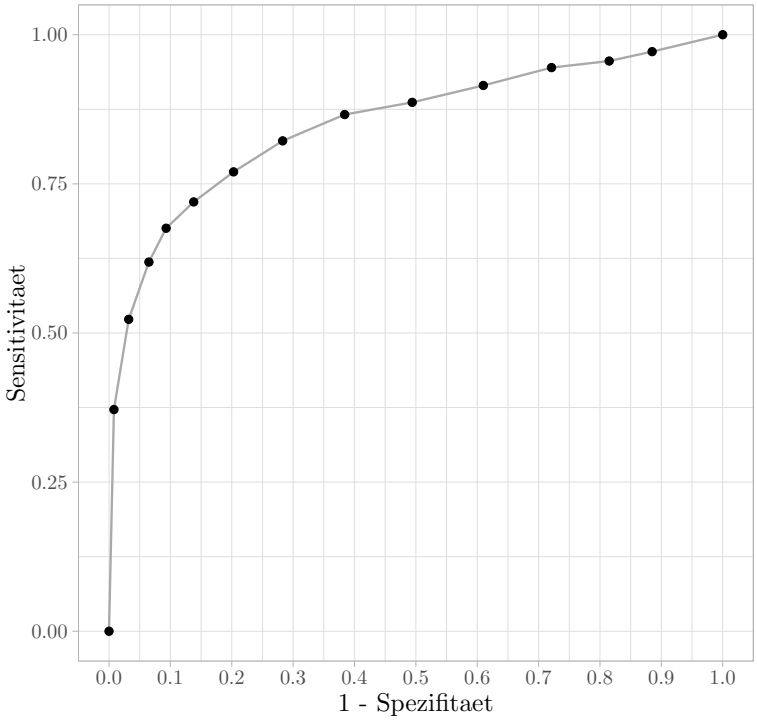

Abbildung 3.7.4: Geschätzte ROC-Kurve aus dem c-log-log Modell

Patienten mittels einer Vordiagnose durchzuführen. Vor allem bei Krankheiten mit einer sehr geringen Prävalenz ist dies sinnvoll.

3.7.3 Optimaler Schwellenwert

Wie in unserem Beispiel bei dem wir kein konkretes Vorwissen haben, würde man sich für den Threshold entscheiden, der die *False-Positive-Fraction* minimiert und die *Sensitivität* maximiert. Hierzu gibt es mehrere Möglichkeiten um einen optimalen Thresholdwert zu ermitteln, einer von diesen ist der *Youden-Index*. Über diesen Index können wir den Schwellenwert ermitteln, welcher uns die optimale Trennung von *Kranken* und *Gesunden* gewährleisten kann. Hierbei berechnen wir

$$YI(t) = Sens(t) + Spez(t) - 1$$

und ermitteln dann über

$$t_{opt} = \max_t \{YI(t)\}$$

den optimalen Thresholdwert (Krzanowski and Hand, 2009). In unserem Anwendungsbeispiel ergibt sich $t_{opt} = 5.9$ mit $Sens(5.9) \approx 0.68$ und $1 - Spez(5.9) \approx 0.10$.

In Abbildung 3.7.5 wird der *Youden-Index* grafisch veranschaulicht und wir können das Optimum daraus ablesen. Hierbei erkennen wir auch, dass es eigentlich zwei potentielle t_{opt} geben könnte. Je nachdem, wie die jeweiligen Schätzer für unsere *Senstivität* und *Spezifität* aussehen, können wir uns in diesem Fall für einen von beiden Schwellenwerten entscheiden. Würde man nur das Maximum der obigen Optimierungsgleichung suchen, so würde man die weiteren möglichen Optima vermutlich nicht bemerken, daher ist es sinnvoll, sich nicht nur das Maximum, sondern einen kleinen Bereich um das Maximum anzuschauen.

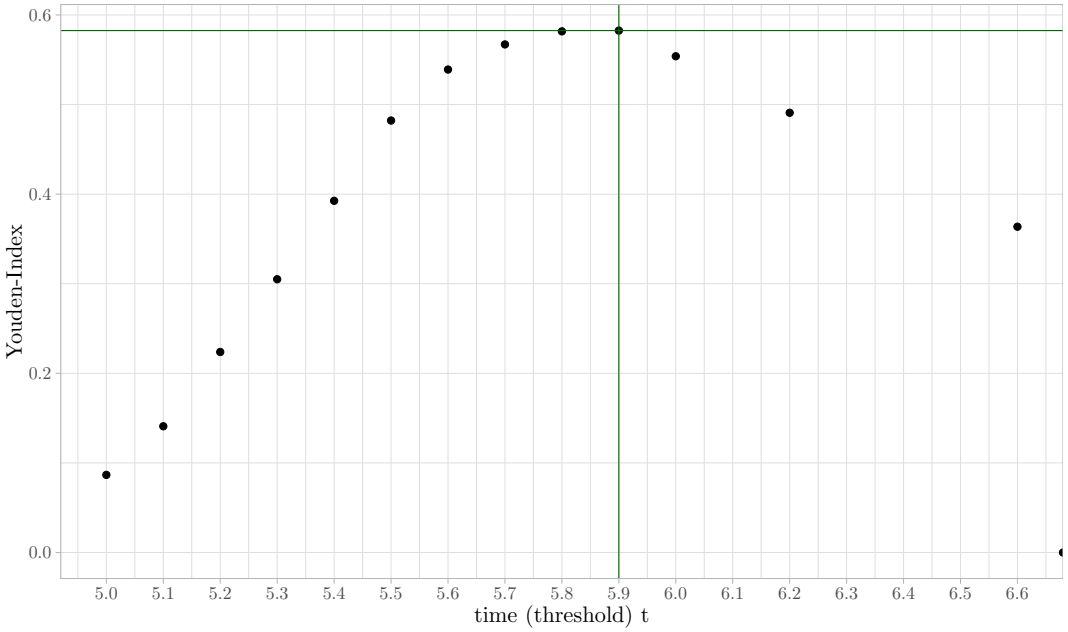

Abbildung 3.7.5: Maximum des Youden-Indexes

4. Visueller Vergleich: ROC-GLM-Schätzer

In diesem Kapitel möchten wir einen ersten visuellen Vergleich beider Modelle - des complementary-log-log Modells und des ROC-GLM-Schätzers - durchführen. Wir bereits im Vorfeld besprochen wurde, werden in den meisten biomedizinischen und epidemiologischen Publikationen keine Individualdaten zur Verfügung gestellt. Stattdessen werden bei Diagnosestudien lediglich speziell ausgewählte Thresholds zur Ermittlung der Sensitivität sowie der Spezifität bereitgestellt. Bisher konnten wir mittels dieser bereitgestellten gruppierten Daten nur eine empirische, also nicht auf ein Regressionsmodell gestützte, ROC-Kurve schätzen, da wir keinen Zugriff auf die Individualdaten haben um damit eine klassische ROC-GLM-Schätzung durchzuführen. Nun greifen wir auf die zuvor eingeführte Idee der intervall-zensierten Daten zurück. Diese Datenaufbereitung hat es uns ermöglicht eine ROC-Kurve mittels des generalisierten Regressionsmodells mit complementary-log-log Linkfunktion zu schätzen ohne dabei auf die Individualdaten zurückzugreifen. Daher möchten wir in diesem Kapitel einen bereits bekannten Datensatz mit Individualdaten aus einer Diagnosestudie hernehmen um zuerst einen klassischen ROC-GLM-Schätzer darauf anzupassen. Als Nächstes werden wir diesen Datensatz geeignet gruppieren um das complementary-log-log Modell auf intervall-zensierte Daten anwenden zu können. Die Ergebnisse beider Modelle werden wir in einer ROC-Kurve darstellen und miteinander visuell vergleichen.

4.1 ROC für Individualdaten

4.1.1 Datengrundlage

Als Beispiel betrachten wir den Datensatz "Pancreatic cancer serum biomarkers study", welcher zum ersten Mal von Wieand et al. (1989) veröffentlicht wurde. Dieser Datensatz beinhaltet die Daten einer Fall-Kontroll-Studie mit 90 Fällen mit Bauchspeicheldrüsenkrebs und 51 Kontrollen, welche lediglich eine Bauchspeicheldrüsenentzündung vorwiesen. Es wurden von jedem Patienten Serumproben (Biomarker y) entnommen und auf einer positiv stetigen Skala wurden das Krebsantigen (cancer antigen) *CA-125* sowie das Kohlenhydratantigen (carbohydrate antigen) *CA-19-9* gemessen. Die Rohdaten `wiedat2b.csv` können über den beigefügten Datenträger vom Anhang A geladen werden.

ID	d	y	test
1	0	28.00	CA-19-9
2	0	15.50	CA-19-9
3	0	8.20	CA-19-9
⋮	⋮	⋮	⋮
140	1	19.2	CA-125
141	1	14.2	CA-125

Tabelle 4.1.1: Datenaufbereitung von `wiedat2b.csv`

In Tabelle 4.1.1 sehen wir einen bereits aufbereiteten Ausschnitt des oben eingeführten Datensatzes `wiedat2b.csv`. Wir werden diesen Datensatz nutzen um eine ROC-Kurve mittels des *ROC-GLM*-Ansatzes zu schätzen. Jedoch beschränken wir uns für die Anschaulichkeit nur auf den Test *CA-125*.

4.1.2 Binormale ROC-Kurve

Der ROC-GLM-Schätzer ist ein moderner Ansatz zur Schätzung einer ROC-Kurve. Dieser Schätzer verwendet den Zusammenhang mit der sogenannten *binormalen ROC-Kurve*. Dafür betrachten wir zuerst folgendes Resultat.

Resultat 3 (Binormale ROC-Kurve)**.**
Seien $Y_K \sim N(\mu_K, \sigma_K^2)$ *und* $Y_G \sim N(\mu_G, \sigma_K^2)$ *dann gilt*

$$ROC(x) = \Phi(a + b\Phi^{-1}(x))$$

mit $x \in (0,1)$ *wobei*

$$a = \frac{\mu_K - \mu_G}{\sigma_K}, \quad b = \frac{\sigma_G}{\sigma_K}$$

und $\Phi(\cdot)$ *die Verteilungsfunktion der Standardnormalverteilung bezeichnet.*

Begründung. Für jeden beliebigen Threshold t gilt:

$$FPF(t) = P[Y_G > t] = \Phi\left(\frac{\mu_G - t}{\sigma_G}\right), \quad TPF(t) = P[Y_K > t] = \Phi\left(\frac{\mu_K - t}{\sigma_K}\right).$$

Für die False-Positive-Fraction c gilt somit, dass $t = \mu_G - \sigma_G\Phi^{-1}(c)$ der entsprechende Threshold für positives Testergebnis. Also gilt

$$\begin{aligned} ROC(c) = TPF(t) &= \Phi\left(\frac{\mu_K - t}{\sigma_K}\right) \\ &= \Phi\left(\frac{\mu_K - \mu_G + \sigma_G\Phi^{-1}(c)}{\sigma_K}\right) \\ &= \Phi(a + b\Phi^{-1}(c)). \end{aligned}$$

□

Definition 10 (binormale ROC-Kurve)**.**
Für $x \in (0,1)$ *definieren wir*

$$ROC(x) = \Phi(a + b\Phi^{-1}(x)).$$

4.1.3 Idee des ROC-GLM-Schätzers

Um einen Schätzer für eine parametrische ROC-Kurve zu erhalten, verwenden wir den ROC-GLM-Schätzer, welcher auf Platzierungswerten basiert. Da bei diesem Schätzverfahren nur die Ränge eine Rolle spielen, ist dieser Ansatz invariant gegenüber strikt monotoner Transformationen. Mit Hilfe von Definition 10 werden die Parameter a und b mittels binärer *Probit-Regression* geschätzt. Es wird keine Verteilungsannahme über die Daten getroffen, nur eine parametrische Annahme über die Form der ROC-Kurve. Für die genaue Schätzgleichung verweise ich an dieser Stelle auf Pepe (2003, Kapitel 5.5).

Auf dem beigefügten Datenträger befindet sich die Datei `ROC_GLM.R` im Anhang A, welche den R-Code für die ROC-GLM-Schätzung beinhaltet. Wenden wir diesen nun auf unsere oben eingeführten Daten an und beschränken wir uns nur auf den Test *CA-125*, so erhalten wir die ROC-Kurve in Abbildung 4.1.1. Die dazugehörigen geschätzten Parameter können ebenfalls mithilfe der Datei `ROC_GLM.R` reproduziert werden.

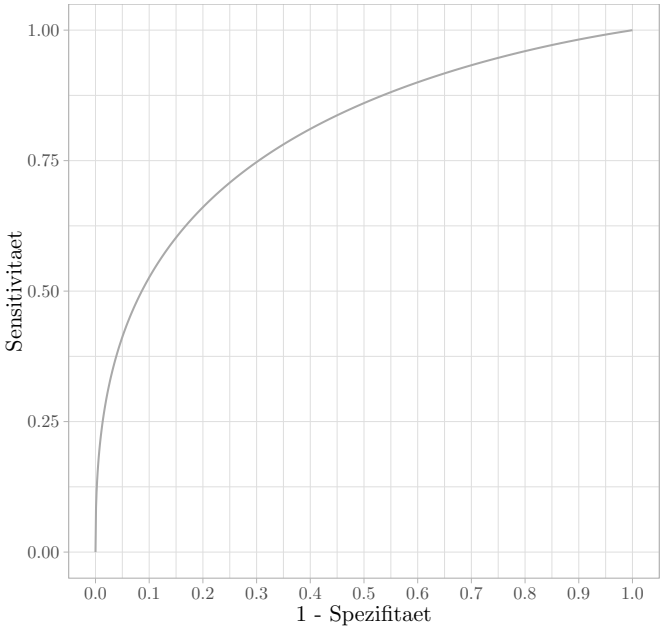

Abbildung 4.1.1: ROC-Kurve über ROC-GLM Schätzer

4.2 ROC für intervall-zensierte Daten

4.2.1 Datenaufbereitung - Gruppierung

Um den ROC-GLM-Schätzer nun mit unserem complementary-log-log Modell zu vergleichen müssen wir den oben eingeführten Datensatz `wiedat2b.csv` in eine geeignete Form bringen um ihn mit den Methoden aus Kapitel 3 analysieren zu können. Hierfür müssen wir unsere Individualdaten in eine gruppierte Datenform überführen. Als Erstes werden wir uns passende Thresholds überlegen. Dafür eignet es sich am besten, bestimmte Quantile auszusuchen und diese als Thresholdwerte festzulegen. Danach müssen wir nur noch die Kennzahlen TPF, TNF, FPF und FNF bestimmen und den dadurch resultierenden Datensatz wie im Abschnitt 3.5 mit den bereits erstellen Funktionen aufbereiten.

Verwenden wir die Daten aus Tabelle 4.1.1. An dieser Stelle müssen wir uns für Thresholdwerte entscheiden, diese wählen wir über die nachfolgenden Quantile bezüglich y aus:

p	0	0.15	0.30	0.45	0.60	0.75	0.90
p-Quantil	3.7	9.1	11.6	15.0	21.4	35.0	79.1

Tabelle 4.2.1: Thresholdwerte

Für jeden einzelnen Thresholdwert aus Tabelle 4.2.1 werden wir eine 4-Felder-Tafel erzeugen, indem wir alle Individuen als *Positive* klassifizieren, die in diesem Fall beispielsweise einen Thresholdwert größer als 3.7 vorweisen. Der Klassifizierungsstatus wird mit dem echten Krankheitsstatus (Variable d) verglichen und darüber werden die *Fractions* ermittelt. Für den Threshold 3.7 führt es beispielhaft zu folgendem Output:

Threshold	TP	TN	FP	FN
3.7	177	3	99	3

Tabelle 4.2.2: Datenauszug - gruppierte Daten

Führen wir dies sukzessive fort und bringen den neuen Datensatz in die Inputform wie im Abschnitt 3.5, so erhalten wir unseren neuen Datensatz in Tabelle 4.2.3. Hierbei sind die Spalten so ausgewählt, dass der Datensatz kompatibel mit den Funktionen aus Abschnitt 3.5 ist. Dabei wurde der Wert von Variable `study=2` gesetzt; dieser Wert ist willkürlich, jedoch verlangt unsere Funktion einen Inputwert für die Variable `study`, da die Funktionen so programmiert wurden, dass man sie für die Meta-Analyse erweitern könnte.

study	TP	TN	K	G	Threshold
2	177	3	180	102	3.7
2	161	38	180	102	9.1
2	152	56	180	102	11.6
2	136	65	180	102	15.0
2	117	77	180	102	21.4
2	95	89	180	102	35.0
2	70	97	180	102	79.1

Tabelle 4.2.3: Inputdatensatz generiert von `wiedat2b.csv`

Nachdem wir unseren Datensatz `wiedat2b.csv` entsprechend gruppiert haben, können wir nun wieder von intervall-zensierten Daten ausgehen und unsere Algorithmen aus Kapitel 3 anwenden, um ein entsprechendes complementary-log-log Modell zu schätzen. In Abbildung 4.2.1 haben wir dies getan und sehen hier unsere ROC-Kurve aus dem complementary-log-log Modell. Die dabei geschätzten Parameter können mithilfe der Datei `ROC_GLM.R` im Anhang A reproduziert werden.

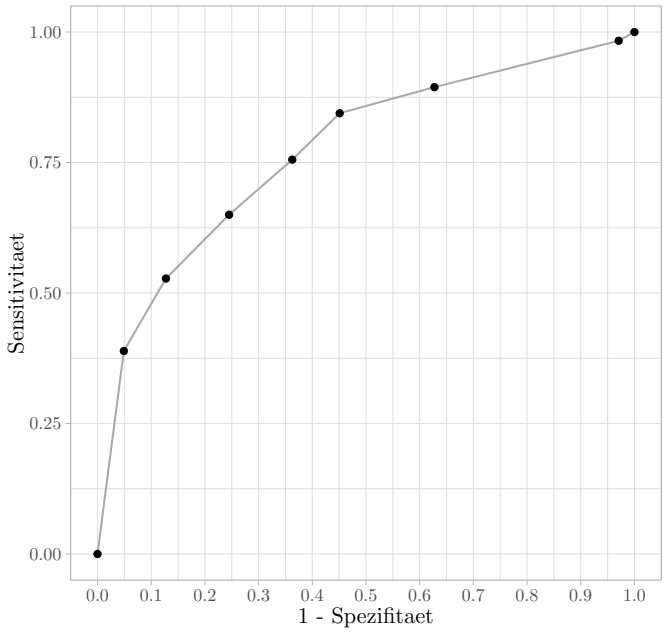

Abbildung 4.2.1: ROC-Kurve über complementary-log-log Modell

Vergleichen wir diese zwei Kurven auf einer visuellen Ebene, so können wir, wie bereits erwartet, eine relativ gute Anpassung der ROC-Kurve des complementary-log-log Modells an den ROC-GLM-Schätzer erkennen. In Abbildung 4.1.1, welche den ROC-GLM-Schätzer verwendet, erhalten wir eine stetige Kurve, bei der wir zu jeder Spezifität auch eine Sen-

sitivität ablesen können. Die dazugehörigen Thresholdwerte kann am anhand des Modelloutputs im Anhang A über die geschätzten Parameter ablesen, so wie es in Tabelle 3.7.1 bereits gemacht wurde. Daher wäre es denkbar, die Thresholds in weiterführende Grafiken zu integrieren. Im Vergleich zu Abbildung 4.2.1, welche über das complementary-log-log Modell geschätzt wurde, wird hier nur die Sensitivität und die Spezifität in Abhängigkeit vom gegebenen Thresholdwert t ermittelt. Da es sich hierbei nur um eine erste und rein visuelle Gegenüberstellung beider Verfahren handelt, wäre es sinnvoll, eine Simulationsstudie durchzuführen, um genauere statistische Aussagen treffen zu können.

5. Diskussion und Ausblick

Das Ziel dieser Abschlussarbeit bestand darin, ROC-Kurven über das complementary-log-log Modell zu schätzen. Es wurde gezeigt, dass wir bei vorliegenden gruppierten Daten diese als intervall-zensiert betrachten können und dadurch Pseudobeobachtungen erzeugen. Nach geeigneter Datenaufbereitung können wir also eine ROC-Kurve mittels eines complementary-log-log Modells parametrisch schätzen. Der meiste Aufwand liegt vor allem in der korrekten Datenaufbereitung, also in der Erzeugung von Pseudobeobachtungen, hierfür wäre es im Weiteren praktisch ein R-Paket mit entsprechenden Funktionen bereitzustellen.

Wie wir beim visuellen Vergleich des complementary-log-log Modells und des ROC-GLM-Schätzers zur Schätzung einer ROC-Kurve gesehen haben, sind die beiden in Form des Graphen aneinander gut angepasst. Nichtsdestotrotz sollte man daran denken, dass wir beim complementary-log-log Modell auf Grundlage von gruppierten bzw. intervall-zensierten Daten schätzen, welche somit weniger Information als Individualdaten in sich tragen. Wie zuvor erwähnt wurde, liegen in wissenschaftlichen Publikationen keine oder nur sehr selten Individualdaten, sondern meistens nur gruppierte, also intervall-zensierte, Daten vor. Mit dieser Informationslage stellt das complementary-log-log Modell eine gute Möglichkeit dar, eine ROC-Kurve parametrisch zu schätzen. Liegen jedoch Individualdaten vor, sollte die ROC-Kurve beispielsweise mittels des ROC-GLM-Schätzers ermittelt werden.

Im Anschluss dieser Arbeit ist es sinnvoll eine Simulationsstudie durchzuführen, in der man beide Modelle bezüglich eines geeigneten Gütekriteriums miteinander vergleicht. Des Weiteren ist es auch sinnvoll sich genauer mit den Konfidenzintervallen auseinanderzusetzen um die Unsicherheit der einzelnen Parameter besser quantifizieren zu können sowie weitere mögliche Kennzahlen wie zum Bespiel der AUC zu ermitteln.

Aufbauend auf dieser Arbeit wäre es denkbar das eingeführte complementary-log-log Modell für intervall-zensierte Daten zur Schätzung von Sensitivität und Spezifität auf mehr als eine Studie zu erweitern und es somit für Meta-Analysen anzupassen. Dies führt uns zu einem generalisierten linearen gemischten Modell mit complementary-log-log Link.

A. Elektronischer Anhang

Als elektronischer Anhang zu dieser Arbeit ist eine CD beigefügt, welche unter anderem die verwendeten Datensätze sowie selbsterstellte Funktionen und Algorithmen enthält. Folgende Dateien sind auf dem Datenträger enthalten:

- `main_data_process.R`
  R-Code, welcher für den Hauptteil der Arbeit verwendet wurde, Erzeugung der aufbereiteten Daten, Anpassung des complementary-log-log Modells, Berechnung der Sensitivität und Spezifität, Generierung der Life-Tables und der ROC-Kurve aus Kapitel 3
- `functions.R`
  Grundlegende Funktionen zur Datenaufbereitung und Generierung von Pseudobeobachtungen
- `ROC_GLM.R`
  R-Code, welcher für die Analyse des ROC-GLM-Schätzers verwendet wurde sowie grafischer Output der ROC-GLM-Kurve und der ROC-Kurve aus dem complementary-log-log Modell aus Kapitel 4
- `BA_Wisskott.pdf`
  PDF-Version der Bachelorarbeit

Bei der Erstellung der R-Codes wurden folgende R-Pakete verwendet:

- `ggplot2` Wickham (2016)
- `tidyverse` Wickham et al. (2019)
- `tikzDevice` Sharpsteen and Bracken (2019)

Selbständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

München, 20.03.2020 .....

Ort, Datum

.....

Unterschrift

Literaturverzeichnis

P. D. Allison. Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13:61–98, 1982. ISSN 00811750, 14679531. URL http://www.jstor.org/stable/270718.

S. H. Choi, T. H. Kim, S. Lim, K. S. Park, H. C. Jang, and N. H. Cho. Hemoglobin a1c as a diagnostic tool for diabetes screening and new-onset diabetes prediction. *Diabetes Care*, 34(4):944–949, 2011. ISSN 0149-5992. doi: 10.2337/dc10-0644. URL https://care.diabetesjournals.org/content/34/4/944.

A. Hoyer, S. Hirt, and O. Kuss. Meta-analysis of full roc curves using bivariate time-to-event models for interval-censored data. *Research Synthesis Methods*, 9, 10 2017. doi: 10.1002/jrsm.1273.

W. J. Krzanowski and D. J. Hand. *ROC Curves for Continuous Data*. Chapman & Hall/CRC, 1st edition, 2009. ISBN 1439800219.

M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Sciences Series, 2003.

C. Sharpsteen and C. Bracken. *tikzDevice: R Graphics Output in LaTeX Format*, 2019. URL https://CRAN.R-project.org/package=tikzDevice. R package version 0.12.3.

G. Tutz and M. Schmid. *Modeling Discrete Time-to-Event Data*. Springer, Cham, 2016.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

S. Wieand, M. H. Gail, B. R. James, and K. L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3):585–592, 1989. ISSN 00063444. URL http://www.jstor.org/stable/2336123.